Университет ИТМО

Практическая работа №3

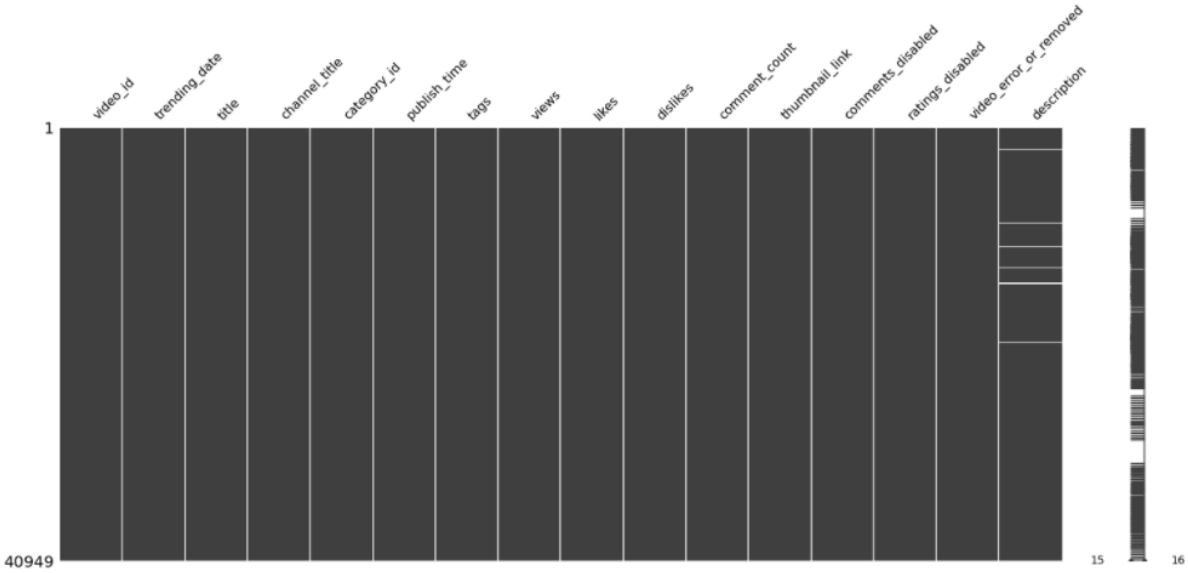
по дисциплине «Визуализация и моделирование»

Автор: Сулейманов Руслан Имранович
Поток: 1.1 **Группа:** к3241 **Факультет:** ИКТ
Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

Посмотри в каких колонках есть пустые значения

```
In [3]: # в каких колонках есть пустые значения
        msno.matrix(df)
Out[3]: <AxesSubplot:>
```



Видем, что у нас хороший датасет, и изменять его фактически не надо. Поэтому только переделаем структуру. Первое что мы сделаем, нормальный лист тегов, а не длинную строку.

```
:
# конвертируем список тегов в list
def to_list(s):
    return list(s.replace('"', '').split('|'))

df.loc[df["tags"] == "[none]", "tags"] = ""
df["tags"] = df["tags"].apply(to_list)
df[["tags"]].head()
```

:

	tags
0	[SHANtell martin]
1	[last week tonight trump presidency, last week...
2	[racist superman, rudy, mancuso, king, bach, r...
3	[rhett and link, gmm, good mythical morning, r...
4	[ryan, higa, higatv, nigahiga, i dare you, idy...

Далее, выбросим ненужные столбец, нам совсем не нужно знать ссылку на видео.Выбросим его: $df = df.drop(columns = [”thumbnail_link”])$

Далее конвертируем дату выхода в тренда и дату публикации из строки в нужный datetime:

```
# конвертируем дату выхода в тренда и дату публикации из строки в datetime
df["trending_date"] = pd.to_datetime(df["trending_date"], format="%y.%d.%m")
df["publish_time"] = pd.to_datetime(df["publish_time"], format="%Y-%m-%dT%H:%M:%S.%fZ")
df[["trending_date", "publish_time"]].head()
```

	trending_date	publish_time
0	2017-11-14	2017-11-13 17:13:01
1	2017-11-14	2017-11-13 07:30:00
2	2017-11-14	2017-11-12 19:05:24
3	2017-11-14	2017-11-13 11:00:04
4	2017-11-14	2017-11-12 18:01:41

Теперь перезапишем категории в читаемый вид, а именно вместо цифр ссылок на категории запишем значение:

```
def rewrite_categories(i):
    return categories[str(i)]["snippet"]["title"]

df["category_id"] = df["category_id"].apply(rewrite_categories)
df = df.rename(columns={"category_id": "category"})
```

Ну и сохраним все в удобном формате для работы с питоном:

```
# сохраняем
df.to_pickle("df.pkl")
df_unique = df.drop_duplicates(subset="video_id", )
df_unique.to_pickle("df_unique.pkl")
```

Теперь составим гипотезы:

1) видео попадают в тренды в среднем за 3 дня - так как старые видео не смогут резко набрать популярность.

2) при большем отношении дизлайков к лайкам выше отношение комментариев к просмотрам - когда зрителю не нравится видео он пойдет писать гневный комментарий.

3) из типов видео больше всего просмотров в среднем набирают музыкальные клипы - так как музыкальные клипы можно повторять, слушать на фоне, и их больше всего раскручивают.

4) у видео с бОльшим количеством тегов быстрее попадает в тренды - чем больше тегов, тем больше шанс у пользователя увидеть видео.

5) на клипах меньше отношение лайков к просмотрам, чем у других видео в среднем - так как их просмотры сильно раздуты от пересматривания